SINGLE CHANNEL SPEECH ENHANCEMENT WITH RESIDUAL LEARNING AND RECURRENT NETWORK

A Master Thesis
Presented to
The Academic Faculty

By

Hua Chen

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in the
School of Electrical and Computer Engineering

Georgia Institute of Technology

May 2020

SINGLE CHANNEL SPEECH ENHANCEMENT WITH RESIDUAL LEARNING AND RECURRENT NETWORK

Approved by:

Dr. David V. Anderson, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Mark A. Davenport
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Chin-hui Lee
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Kwan Truong
Poly Media Labs
Plantronics

Date Approved: April 21, 2020

To my parents

ACKNOWLEDGEMENTS

I would like to thank Dr. David V. Anderson for this wonderful thesis opportunity as well as his guidance and support. Thanks to everyone in lab ESP who advised and helped me along the way. To Dr. Chin-hui Lee's, his lecturing on Digital Speech Processing inspired my interest in speech processing; To Dr. Mark A. Davenport, his materials in Advanced Signal Processing greatly enhanced my knowledge.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

For speech enhancement tasks, non-stationary noise such as babble noise is much harder to suppress than stationary noise. In low SNR environment, it is even more challenging to remove noise without creating significant artifacts and distortion. Moreover, many state-of-the-art deep learning based algorithms propose a multiple time-frames to one time-frame regression model. In our work, we propose a speech de-noising neural network adopting multiple time-frames to multiple time-frames approach, aiming to greatly reduce computation burden for real-world applications as well as maintain decent speech quality.

In this paper, we propose two neural networks, namely ResSE and ResCRN. ResSE takes form of a ResNet architecture and is inspired by DuCNN, an image enhancement network. With its rich and deep structure and the help of residual connections, ResSE is very efficient at extracting spatial-features and is able to outperform traditional log-MMSE algorithms. ResCRN,with the addition of LSTM layers, is capable at both spatial and temporal modeling. It utilizes both local and global contextual structure information and improves speech quality even when faced with unseen speaker and unseen noises, proving that ResCRN is able to generalize quite well.

CHAPTER 1
INTRODUCTION

1.1 Background on Speech Enhancemeent

Traditionally, speech enhancement is the task of improving noisy speech's perceptual qualities and intelligibility based upon human auditory system and linguistic understanding. Background noises can heavily affect the performance and efficiency of numerous important applications such as real-time communications, long-distance conferences, hearing-aids and speech/speaker recognition.

Two major variables for speech enhancement applications are the number of microphones and types of interference. Microphone arrays would generally be quite helpful to de-noise audio: the more microphones we have to provide different contextual audio information, the better the enhanced audio sample[1]. Interference could either be more noise-like such as engine noise and fan noise or speech-like such as babble noise. Respectively, we can classify the former as stationary noise and the latter as non-stationary noise. In this paper, we focus on the single channel speech enhancement problem for both stationary and non-stationary interference.

Speech enhancement has been the focus of research and product commercialization for decades. In general, there are four categories of algorithms to solve the speech enhancement problem. The best known method is the spectral subtraction algorithm, which estimates noise spectrum from the silent/unvoiced frames, and is spectrally subtracted from the noisy speech to reproduce the estimated clean speech. Spectral subtraction was first proposed by Weiss in 1974 [2] through correlation computation and later by Boll [3] in frequency domain. As simple as the method may seem, it laid out the foundational assumption of speech enhancement – that noise is additive.

The second method is the family of statistical model estimation algorithms, including The classic Wiener filtering [4, 5], log-MMSE speech enhancement algorithms [6] and even Kalman Filter [7]. The statistical techniques estimate the Fourier Transform coefficients of the clean signal given the coefficients of the noisy speech. Nevertheless, the estimated clean speech processed by both spectral subtraction and statistical methods all suffer from the notorious musical noise and distortion. The speech enhancement performance is also not guaranteed with non-stationary interference.

The third technique is the decomposition algorithm, which relies on linear algebra principle that the clean speech is hidden in a subspace of the noisy Euclidean space, including Singular Value Decomposition [8] or Principle Component Analysis [9]. Recently non-negative matrix factorization method [10] was proposed and demonstrated superior results. Nevertheless, these methods are limited by the noise model capacity.

Thanks to the rapid development in GPUs that introduces parallel computing with significantly more computation powers, deep learning has shown great success in speech signal processing community such as Automatic Speech Recognition (ASR), audio event recognition and speaker verification/identification. Neural network based method has also shown substantial improvement over traditional methods regards to speech perceptual quality.

There are two general approaches for deep learning based speech enhancement. Inspired by the time-frequency (T-F) masking in speech separation applications, the first approach is a supervised learning algorithm where a deep neural network estimates an ideal binary mask (IBM). The IBM is used to classify a T-F bin as either noise dominant or speech dominant. More recent development replaces the ideal binary mask with the ideal ratio mask, which learns what percentage of a particular T-F bin is speech. Nevertheless, mask-based algorithms fails to fully utilize the rich acoustic context information of spectra. We adopt the second regression approach in this paper, which aims to directly estimate clean speech spectra given a noisy input. Although, deep learning is capable of de-noising

both stationary and non-stationary interference, removing babble noise is still quite challenging. Limited noise samples could also hinder the performance with noise outside of training set. Especially, the reconstruction of clean speech signal in low SNR cases could lead to devastating.

1.2 Literature Review on Deep Learning based Speech Enhancement

In 2013, Xu [11] was among the first to experiment deep learning possibilities on speech enhancement with an regression approach. In their work [11, 12], they laid out the foundation for DNN speech enhancement system (Fig 2.2) using features from log-power spectrum (LPS) as network input and reconstruct the clean speech with noisy phase and inverse Short-Time-Fourier-Transform. One important contribution by Xu is forming a context window by concatenating noisy neighboring LPS frame with the current frame as network input feature. The context window provides rich acoustic information for the network to learn, and therefore better estimate the current clean LPS frame. The system consists of three fully connected layers, each with 2048 nodes. Although the network design is simple, the results overpowered traditional methods by a large margin in both perceptual and distortion evaluations. The results were further improved by global variance equalization, drop-out training and noise-aware training which feeds the network with feature vectors of estimated noise information.

Since then, the speech community has been quite active to apply state-of-the-art deep learning algorithms for speech de-noising applications. Tu and Zhang [13] proposed to apply skip-connections on feed-forward neural networks. In contrast to most typical ResNet-like networks, the authors applied feed-forward layers with skip connections and constructed residual blocks similar to that proposed by He[14]. Tu and Zhang experimented on learning a log-compressed Mel-Frequency masking using a context window of both past and future frames to predict the clean current frame.

In 2016, Park and Lee [15] proposed two convolutional neural network for speech en-

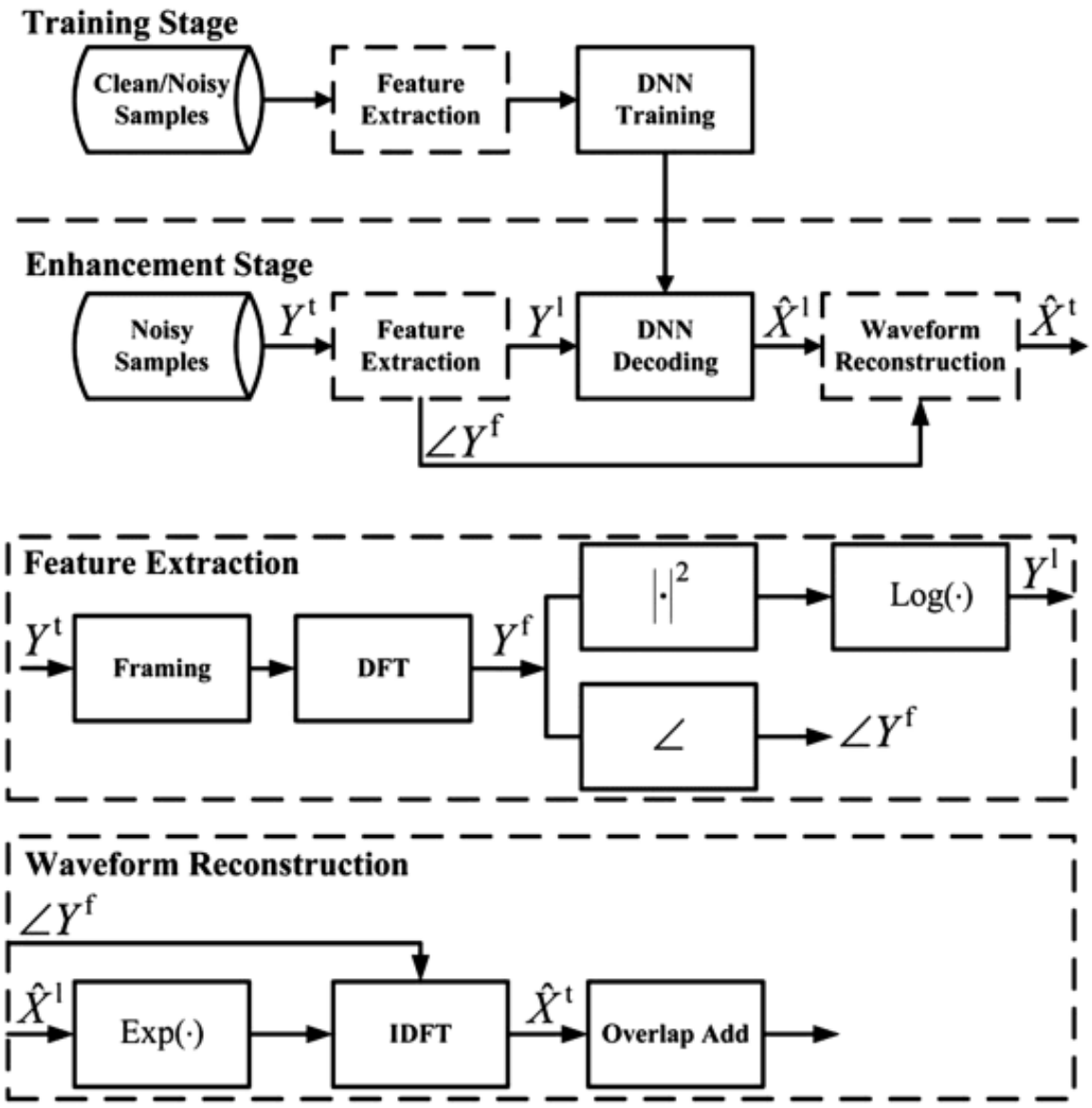


Figure 1.1: DNN based Speech Enhancement system proposed by Xu

hancement. They modified the Convolutional Encoder-Decoder network by adding skip connections between symmetrical layers (CED), and also proposed the Redundant CED network (R-CED). CED encodes the feature into lower dimensions that is later extracted by the decoder; where as R-CED encodes features into higher dimension which is later compressed by the decoder. In their work, the network maps a noisy context window of eight frames to one clean speech frame and pointed out that convolution performed only along the frequency axis yields better results as opposed to the more common 2d convolu-

tions. The proposed R-CED network was shown to restore the clean speech with much less distortion even in low signal-to-noise ratio scenarios. The authors compared thoroughly CED, R-CED and cascaded R-CED with feed-forward neural networks and recurrent neural networks (RNN) and argued that CNN is better suited for speech enhancement tasks while having 100 times smaller parameters.

Tan and Wang [16] proposed an end-to-end convolutional recurrent neural network for speech enhancement. The system is similar to the CED structure introduced above with two LSTM layers between the encoder and decoder structure to exploit long-term context. The model has much fewer parameters compared to its LSTM baseline counterparts and are much better at leveraging the time-frequency structure in magnitude spectrum, leading to significantly better results when dealing with unseen speaker and unseen noises.

Han [17] proposed EHNet, a similar structure to the convolutional recurrent neural network in [16]. EHNet adopts an aggressive filter size of convolution kernels for spectral feature extraction, and proposed two Bidirectional-LSTM layers for temporal modeling. Bidirectional-LSTM layers propagate both forwards and backward and has a better capability of modeling temporal contexts. The author argues that EHNet models clean speech spectrum much better than recurrent network and deep neural network with rich details.

CHAPTER 2
TECHNICAL APPROACH

Speech enhancement problem can be viewed as a multivariate regression problem, mapping a noisy magnitude spectrum to its clean counterpart. It could also be viewed as a filter in complex domain, much like a Gaussian filter for an image. In this section, we discuss in detail our two proposed design: ResSE and ResCRN as well as their intuition. ResSE addresses the speech enhancement problem as an image enhancement task, where the continuous frames in frequency domain are concatenated together to form an image. ResCRN addresses the problem as a many-to-many regression problem. We built ResSE mainly for studying the behaviors of residual connections in the context of speech enhancement and a baseline for comparison with ResCRN.

2.1 Intuition

The intuition for our particular approach is twofold. The first stems from image enhancement work by [18], which aims to enhance degraded image resolution with unknown Gaussian noise through residual connection. The architecture in Fig 2.1 serves as the foundation of our residual learning counterpart for speech enhancement. Note that we can treat spectrograms as images, and thus the speech enhancement problem morphs into an image enhancement problem. Although, spectrum and images are both positive definite, there are three key differences: number of channels, correlations with neighboring pixels and variance. A spectrogram is single channel, whereas images have typically three color channels. Furthermore, while image pixels are highly correlated with all of their neighbors representing textures features, spectrum's pixels correlation are more obscure. Spectrogram time-frequency bin values are not distributed in a certain range; the bins are often depending on the particular window type and sizes. For images, researchers usually normalize the pixel

values from [0,255] down to [0,1]. Lastly, speech enhancement presents more challenging scenarios with babble noise or non-stationary noise as opposed to Gaussian noise in image de-noising applications. However, both could benefit from the idea of residual connection, since in both tasks, the clean signals and degraded outputs are highly correlated. We believe that the residual connections could provide the network with useful information regarding the speech signal.

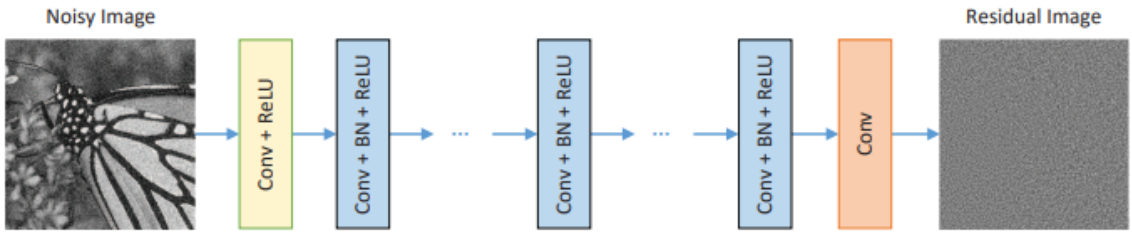


Figure 2.1: Image Gaussian De-noiser Network Architecture

The second intuition is followed by the temporal nature of speech. Speech signals could be viewed as a time series, and its sequential nature is also preserved in frequency domain. While treating spectrogram as an image exploits convolutional neural network's potential, LSTM layers could further improve the results with temporal modeling.

2.2 Proposed Framework

2.2.1 Problem Formulation

We formulate the problem of speech enhancement in our work as finding a mapping between $x \in R_+^{d \times t}$ and $y \in R_+^{d \times t}$, where x and y being the noisy spectrum and clean spectrum respectively with d frequency bins and t time frames. Throughout our work, t is set to 8, specifically 8 consecutive spectrum frames. The mapping function $f(x)$ is learned by neural network on N pairs of noisy and clean spectrum's with the loss function:

$$\min \sum_{n=1}^{N} \|y_n - f(x_n)\|_2^2. \tag{2.1}$$

2.2.2 Residual Network (ResSE)

Inspired by image enhancement work (DnCNN) [18] and ResNet [14], we first design a model to emulate that of the image denoising framework and ResNet framework called ResSE.

The ResSE design could be broken into 5 stages. The first stage is a simple Convolution layer, with 16 filters and a kernel size of $(7, 3)$. Note that we increase the dimension of the kernel along the frequency axis to 7 as opposed to $(3, 3)$ kernels in DnCNN and Resnet. By expanding the receptive field, the network could better cope with the non-stationary nature of the background noises. Batch normalization and ReLU activation function are used after to ensure more stable and better convergence. No max-pooling layers are used since we do not wish to lose any feature and before every convolution layers in ResSE, the input is zero-padded to ensure the dimensions of the spectrogram is constant. The following three stages is a repetition of convolution block followed by two residual blocks.

Convolution Blocks & Residual Blocks

The convolution block is intended to address the increase in channels between the three stages. Instead of a bottle-neck structure, which is intended for very deep ResNet, we adopt the the basic building block structure shown on the left in Fig 2.2. The kernel size in both convolution blocks and residual blocks are increased to $(7, 3)$. The shortcut is also added with another convolution layer, allowing the residual connection information to match in dimension with that of the main path. Furthermore, the network is also able to make adjustments to the short-cut information flow.

The following two residual blocks is kept with constant channel within that same stage and each has two layers of convolution filters. In other words, the convolution block expands the feature maps while the residual block learns rich context of the feature mapping. The shortcut connections in both blocks provide a choice for the network to either add to or subtract from the current output as needed. Within the residual blocks and convolution

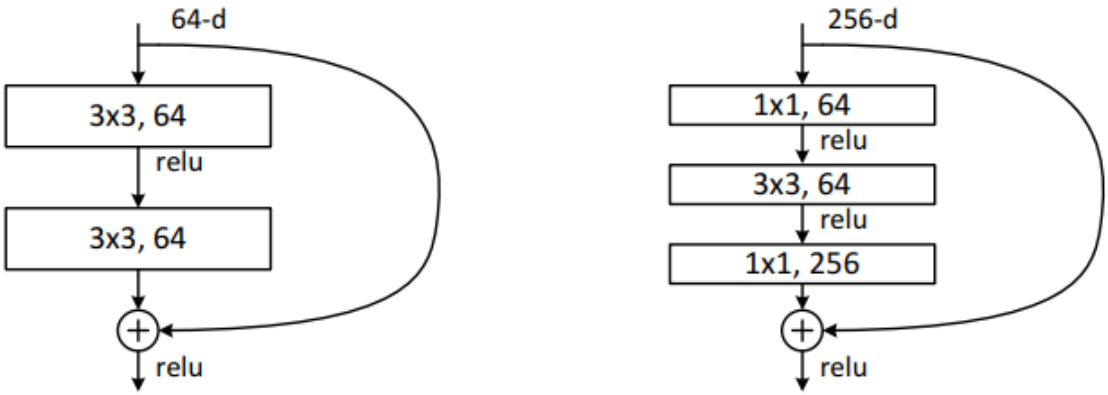


Figure 2.2: ResNet building blocks. The building block on the left is used for shallower ResNet-34, the bottle-neck structure on the right is used for ResNet-50/101/152

blocks, the short-cut is connected after batch normalization and before ReLU activation. The three stages each have $[32\ \ 64\ \ 128]$ feature maps.

The final stage is a single channel convolution filter to reconstruct the clean spectrum. A final ReLU function is used to ensure the non-negative nature of the clean spectra. All convolutions performed in ResSE are 2-dimensional, meaning that we are only learning a spatial de-noising filter for speech enhancement in the frequency domain. Fig 2.3 illustrates our structure at a high level, and Fig 2.4 and 2.5 detail the architecture of convolution blocks and residual blocks.

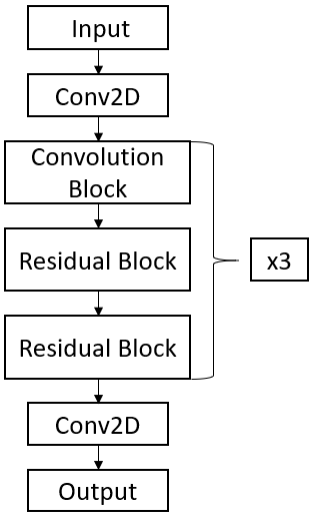


Figure 2.3: ResSE High level Architecture

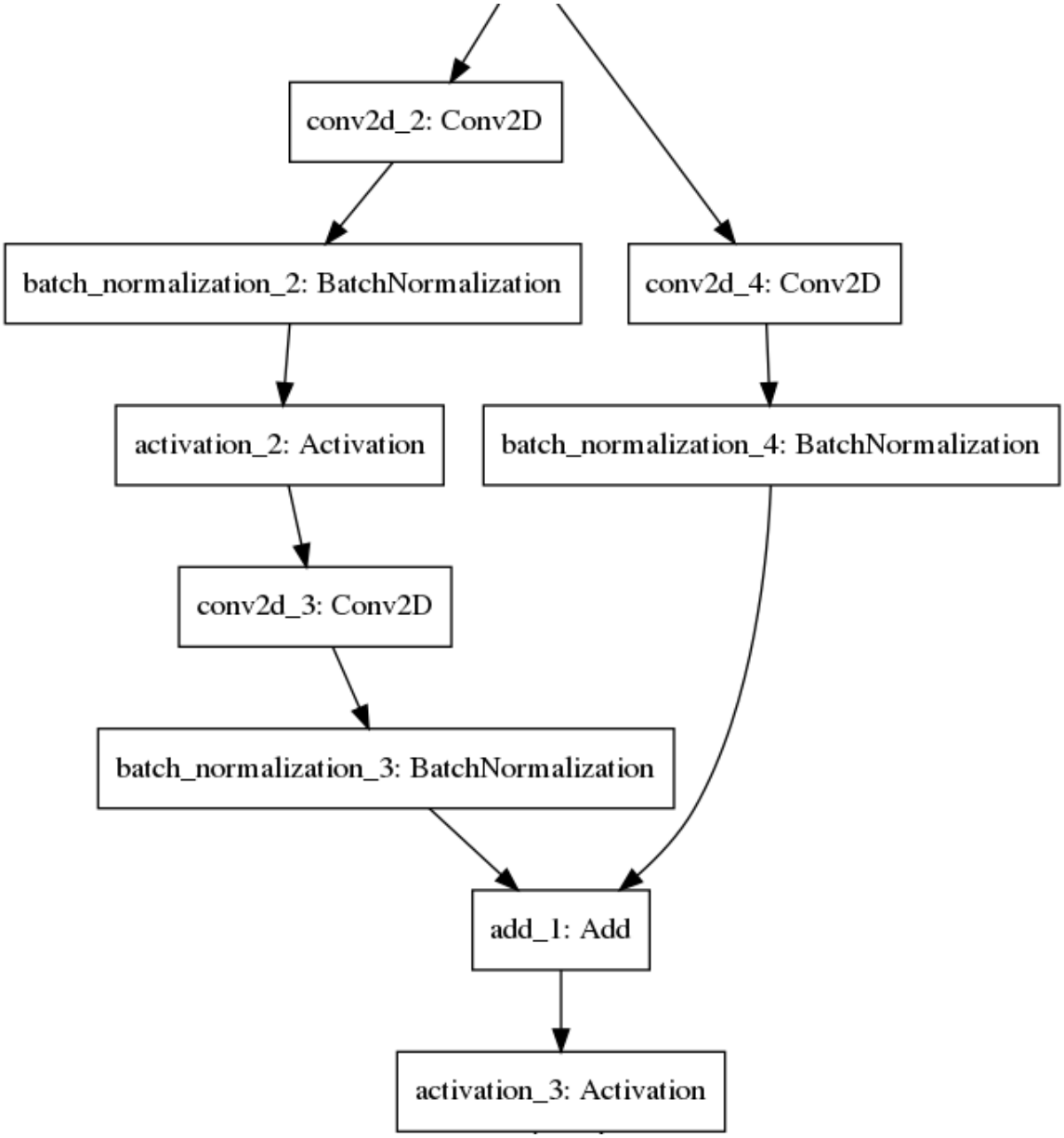


Figure 2.4: Illustration of Convolution Block

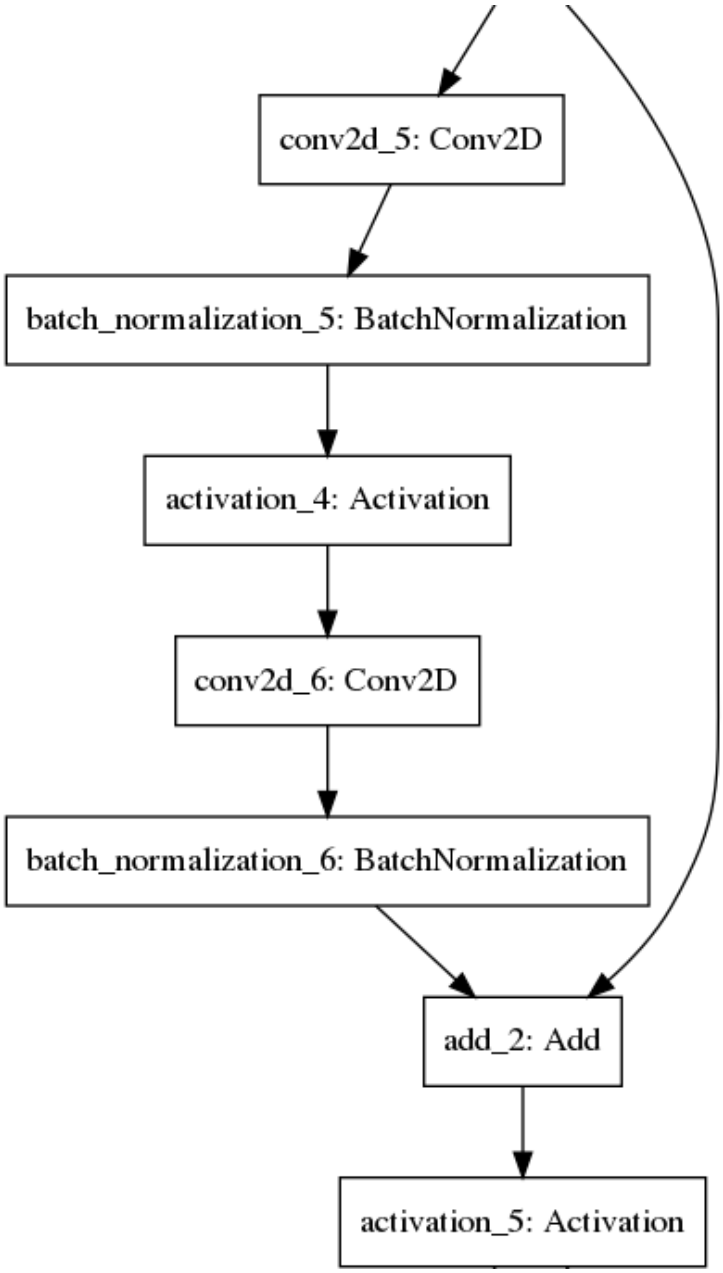


Figure 2.5: Illustration of Residual Block

2.2.3 Residual-based Convolutional Recurrent Neural Network (ResCRN)

While convolution filters are very powerful at learning T-F features in magnitude spectrum, they do not exploit the temporal nature of speech signals. We propose that after the ResSE architecture, we append Long Short-term Memory (LSTM) layers to perform temporal modeling since LSTM layers are able to leverage long-term contexts within our context window. The ResCRN could be broken down into two stages: frequency modeling stage and temporal modeling stage.

In the frequency modeling stage, the convolution operations are used to extract features which are then concatenated and fed to the recurrent neural network. Since recurrent layers bring in a large number of hidden nodes, we wish to keep convolution layers simple yet powerful to alleviate the computation requirements. EHNet [17] has only one convolution layer with 256 kernels and the CRN network proposed in [16] uses five convolution layers of very small filter sizes. Therefore, we decide to keep our frequency modeling stage small but efficient. The first layer performs an aggressive down-sampling with 128 kernels of size $(16, 3)$ and stride $(9, 1)$. The time order is essential for the recurrent network, so the shape is kept the same. After the first layer a residual block same as in ResSE structure is implemented to enrich our feature maps with speech structures. The feature maps are concatenated along frequency axis for the next stage.

In temporal modeling stage, two LSTM layers of 1024 hidden units are used. The

LSTM layers are defined as the following:

$$
\begin{aligned}
i_t &= \sigma(W_{ii}x_t + b_{ii} + W_{h_i}h_{t-1} + b_{h_i}),\\
f_t &= \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}),\\
g_t &= \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}),\\
o_t &= \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}),\\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t,\\
h_t &= o_t \odot \tanh(c_t),
\end{aligned}
$$

where x_t represent the input, g_t is the block input, h_t is the memory cell, the footnote t represents time stamp, W and b are weights and biases, σ is the sigmoid activation function, and $\odot$ represents element-wise multiplication. The final layer is a fully-connected linear regression layer with ReLU function to reconstruct and ensure the non-negative nature of the spectrum.

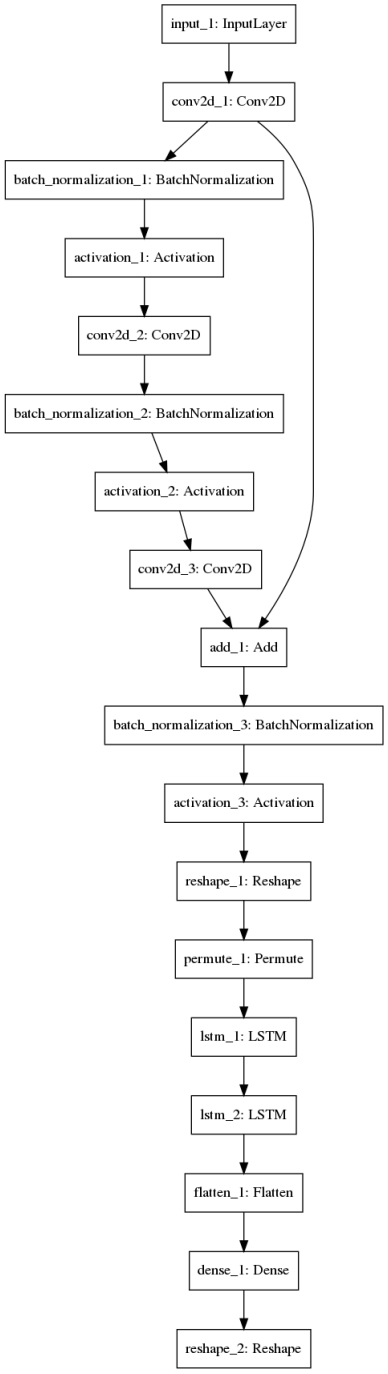


Figure 2.6: ResCRN Architecture

CHAPTER 3

EXPERIMENTS, ANALYSIS AND EVALUATIONS

3.1 Experimental Setup

In the following sets of experiments, we train and evaluate our models on the TIMIT dataset [19]. To cope with our limited computation capability, we build a smaller test set from TIMIT. We randomly selected 100 male speakers and 100 female speakers from the 436 training speakers, each with 10 utterances. We used [20] 100 non-speech environmental sounds as the noise corpus. The training utterances are corrupted with all 100 noises at three levels of SNR, i.e 0dB, 5dB, 10dB. The mixture process results in a 500hr collection of noisy speech data. Since our mixture process closely resembles to that implemented in [12], we adopt the same random selection process to build the training set. The authors pointed out that the network performance saturated between 100 hours to 625 hours randomly selected noisy speech out of 2500 hours of total mixture data, therefore we randomly select 50 hours of our mixture to train both ResSE and ResCRN.

For testing, we evaluate our model's capability with unseen speakers and unseen noise mixtures. 100 utterances were randomly selected from the TIMIT test set. As for unseen noises, we choose three challenging non-stationary sound clips from sound-bible.com [21], namely, airplane noise (an airplane flying by), babble noises (group of people murmuring in the background), and restaurant noises (group of people talking in the background). The testing speech segment are also mixed at SNR = 0dB, SNR = 5dB and SNR = 10dB. All training and testing speech as well as noises are sampled at 16kHz. The frame length for Short-time Fourier Transform is 320 samples (20 msec) and the frame shift is 160 samples (10 msec). The standard single sided spectrum magnitude feature is chosen as the input to our neural network with 161 frequency bins and a context window of 8 frames.

Both models are trained with the Adam optimizer with an initial learning rate of 0.0005, and batch size of 128. ResSE is trained with 40 epochs and ResCRN is trained with 30 epochs. No dropout, or spatial dropout are adopted, since they hurt the performance. The learning rate will be set to decay by a factor of 0.1 if the validation loss does not decrease by 2 epochs. Only the best model with the lowest loss on validation set is saved.

To evaluate the performance, we adopted three main measurements: Perceptual Evaluation of Speech (PESQ) [22], Short-Time Objective Intelligibility (STOI) [23] and Signal to Distortion ratio (SDR) [24]. PESQ scores range from - 0.5 to 4.5 and reflects a high correlation with subjective evaluations. STOI ranges from 0 to 1 in percentage; the higher the score, the better the intelligibility the audio is. All three measurements are calculated against the clean speech. The log-MMSE [6] algorithm is adopted to provide a baseline for our models.

3.2 Experimental Results

SNR(dB)	Noisy	log-MMSE	ResSE	ResCRN
0	0.188	3.583	4.883	**7.799**
5	5.156	8.034	9.715	**12.061**
10	10.146	12.252	13.749	**15.645**

Table 3.1: SDR values for our proposed experiments across 3 unseen noises

SNR(dB)	Noisy	log-MMSE	ResSE	ResCRN
0	0.657	0.638	0.726	**0.784**
5	0.765	0.753	0.819	**0.865**
10	0.857	0.849	0.894	**0.919**

Table 3.2: STOI scores for our proposed experiments across 3 unseen noises

SNR(dB)	Noisy	log-MMSE	ResSE	ResCRN
0	1.383	1.510	1.679	**1.895**
5	1.613	1.801	2.039	**2.246**
10	1.946	2.183	2.329	**2.577**

Table 3.3: PESQ scores for our proposed experiments across 3 unseen noises

In tables 3.1, 3.2, and 3.3, we showcase the results in regard to our evaluations matrix. Log-MMSE method struggles when dealing with highly non-stationary noises, with an average increase of 2.793 in SDR, a 0.013 decrease in STOI, and a 0.184 increase in PESQ. ResSE performs better than traditional signal processing algorithms, due to the rich structure of its design. ResCRN, which models both temporal and frequency features yields much better results than ResSE across all three criteria.

To better understand our performance, we examine the spectrograms with respect to the three unseen noises. The spectrums of noisy, de-noised and clean audio samples are shown below for examination. ResSE is able to remove some level of noises. However, the speech signals are still contaminated especially in the low frequency bins. Due to the small context-window and absence of temporal modeling, ResSE is not very good at differentiating noise

and speech. Nevertheless, the speech structure is still intact with residual connections, hence the distortion is mainly caused by noise residuals. However, ResSE still outperforms log-MMSE in STOI and PESQ.

The ResCRN network results in a better trade-off between removing noise and preserving speech characteristics even with the two challenging babble noise test cases. For the less challenging airplane noise at SNR = 0dB, ResCRN performs noticeably better with very little residual noises. In less noisy scenarios for airplane noise, ResCRN almost fully restored the clean lower frequency T-F bins but there are more distortions in higher frequency range. The same observation could also be found in the two babble noises. For both restaurant and babble noise, ResCRN enhances the lower frequency range bins much better than higher range ones. Since most of the speech spectrum content resides below 8kHz, distortion above 8kHz is less objectionable to human perception.

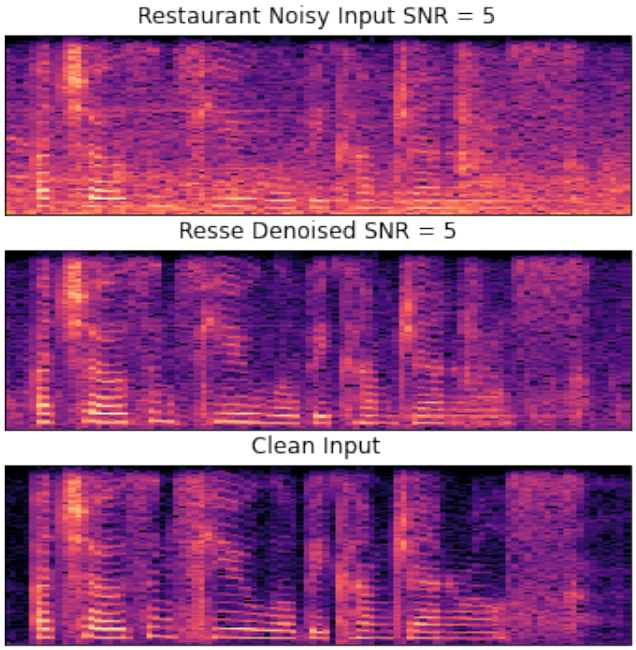


Figure 3.1: Spectrogram of ResCRN denoised

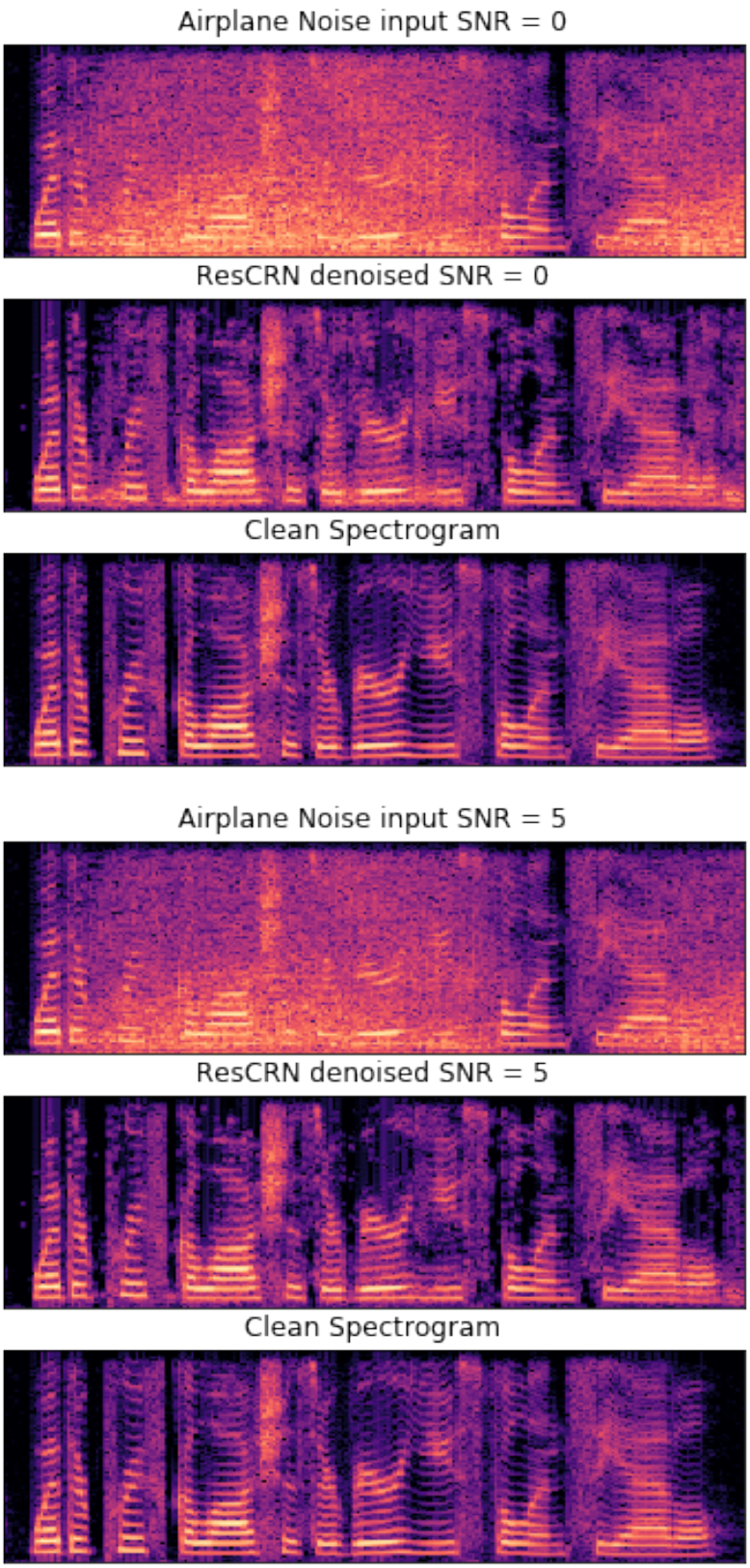


Figure 3.2: ResCRN performance with unseen speaker and unseen noise:Airplane

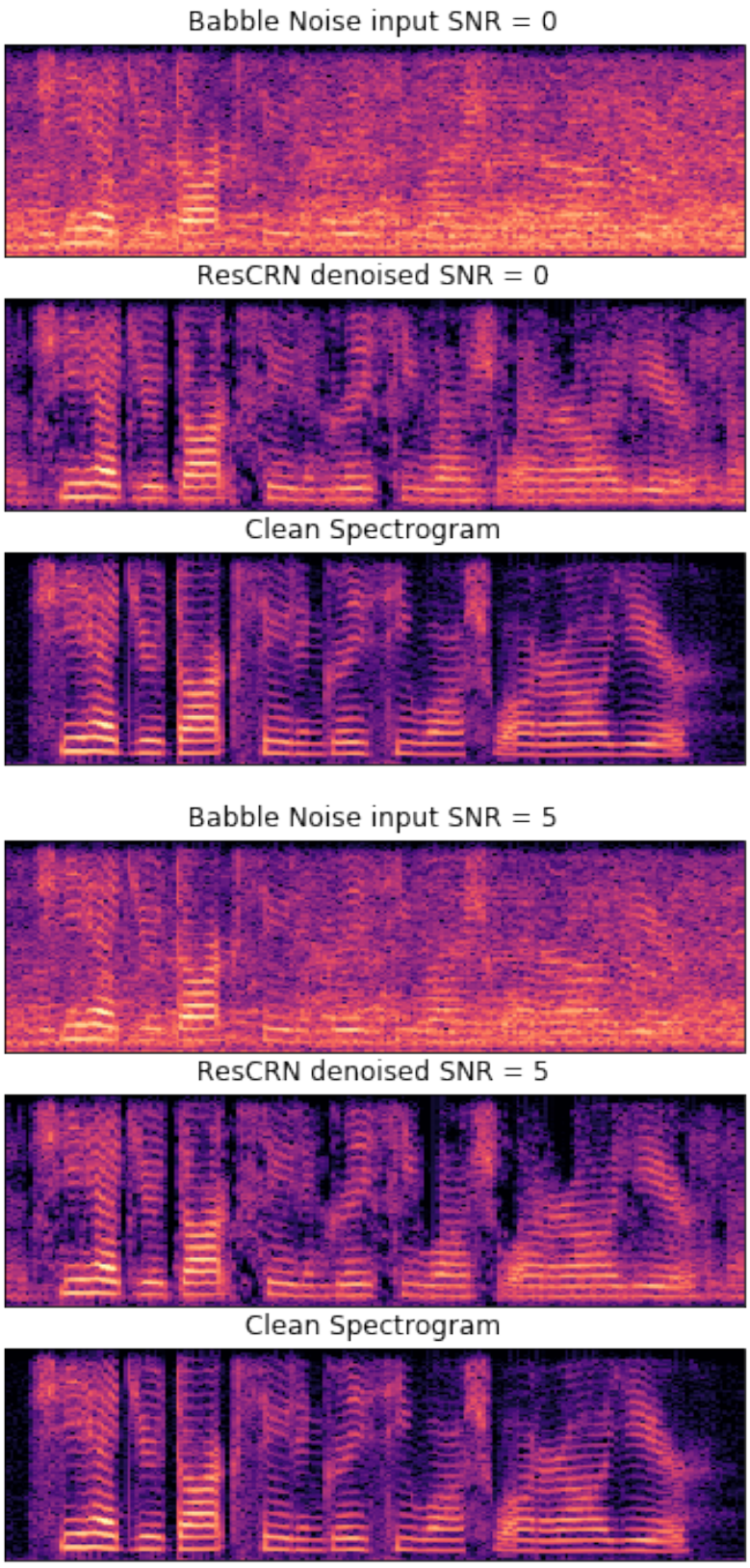


Figure 3.3: ResCRN performance with unseen speaker and unseen noise:Babble

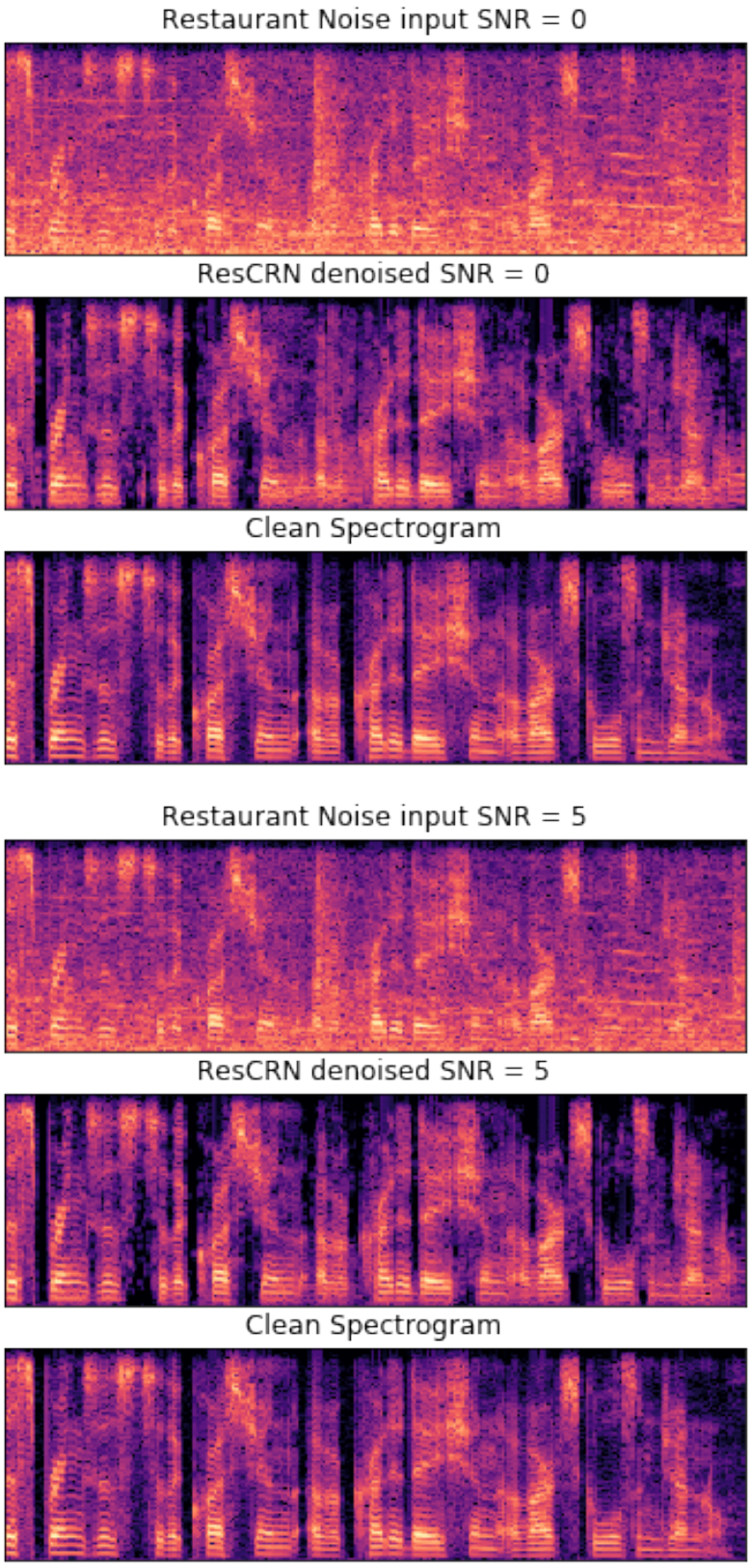


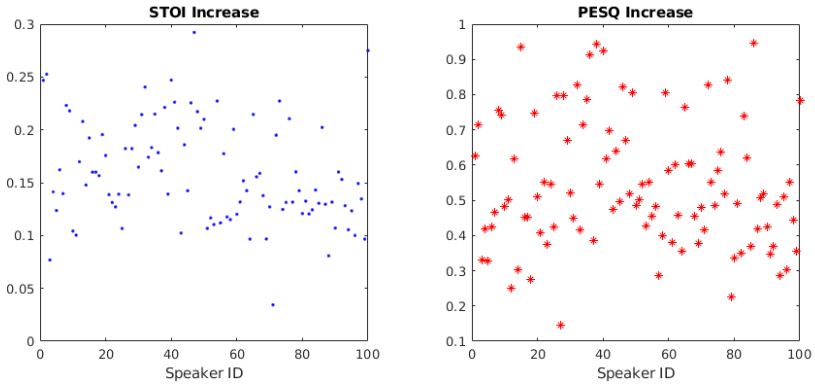
Figure 3.4: ResCRN performance with unseen speaker and unseen noise:Restaurant

CHAPTER 4
DISCUSSION

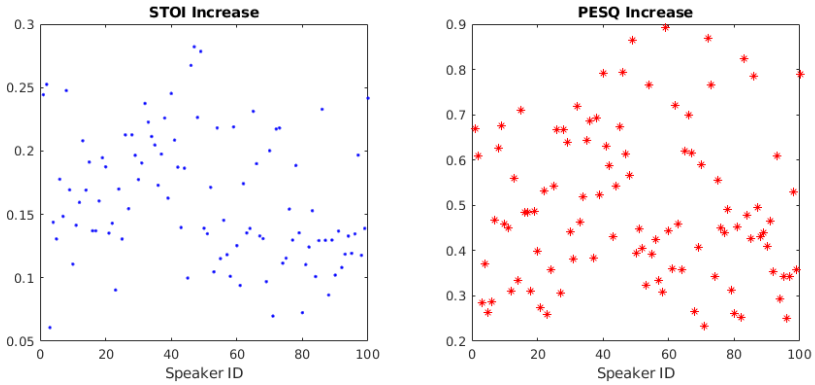
We showed in the previous sections, that our proposed ResSE network and ResCRN network can consistently outperform log-MMSE approach. The noise suppression capability for non-stationary noises is far better than traditional signal processing methods. In this section we discuss lessons we learned during the research, and compare our approach to prior studies.

4.1 Performance across different speakers

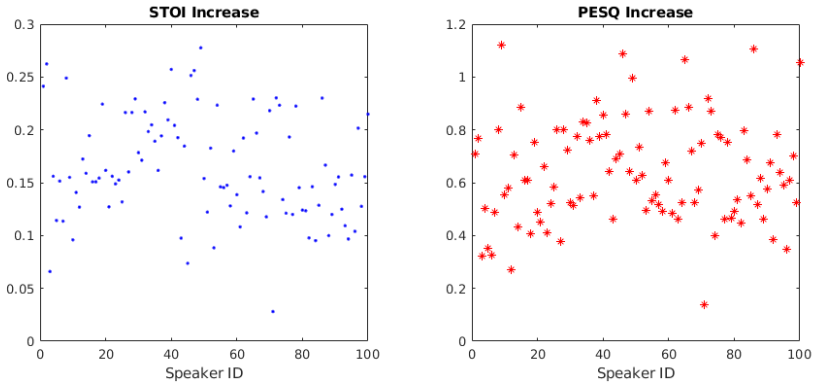
Although DNN based speech enhancement algorithms are able to perform well with unseen speakers and unseen noise types, the performance for each individual speaker is not uniformly distributed. Fig 4.1 illustrates that at SNR = 0dB, the exact STOI and PESQ values increased by ResCRN. The plots are computed by subtracting noisy utterance PESQ and STOI scores from the ResCRN enhanced counterparts. The difference between enhanced and noisy is shown for each speaker. The first 50 speakers are female, and the last 50 speakers are male. For airplane noise, the noise power mostly resides in lower frequency range. It has a bigger impact on male speakers, since female speakers tend to see a bigger increase in both STOI and PESQ scores. Babble noise is the most challenging noise type, and the increase for both male and female speakers are highly variant. Restaurant noise has a more balanced performance boost between the two genders. It is also interesting to see that within each gender, the boost is still quite irregular. This phenomena could be due to TIMIT's broad range of speaker traits and distinct accents.



(a) Airplane Noise



(b) Babble noise



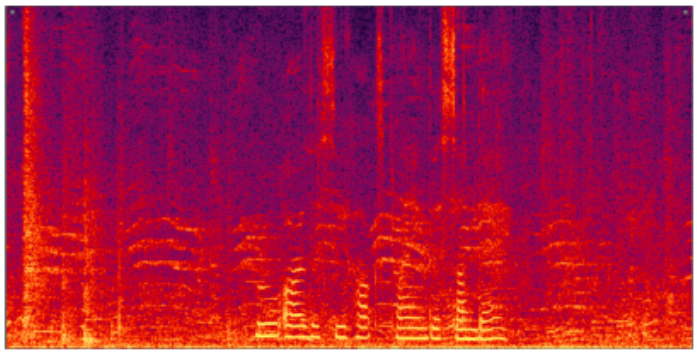
(c) Restaurant Noise

Figure 4.1: STOI and PESQ increase at SNR = 0 for each noise type. The first half of the 100 speakers are female, the second half of speakers are male.

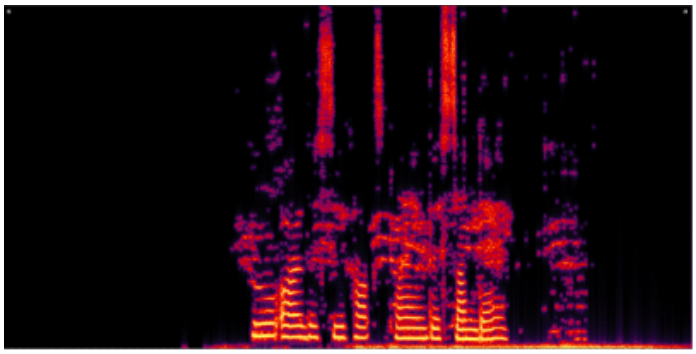
4.2 Distortions

As shown in the figures 3.1 - 3.4, although the speech qualities are greatly improved by our proposed designs, there still are some distortions especially in the higher T-F bins. One of the reasons is that the spectrum values are much lower in higher frequency bins compared to lower ones, making them susceptible to be characterized as noise. Another potential reason is due to noisy phases, especially in heavily corrupted audio sequences. Fortunately, such distortions do not affect perceptual qualities as well as speech intelligibility scores too much.

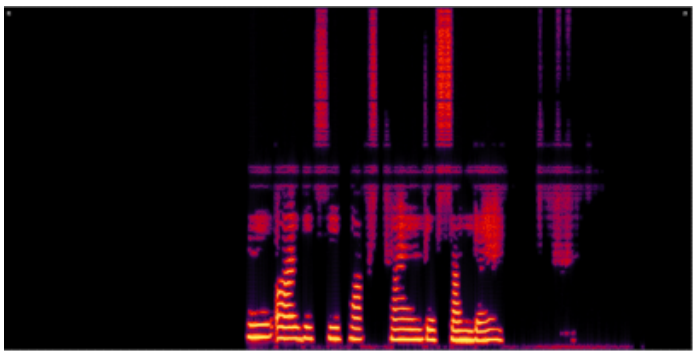
Distortions are common in almost all neural network approaches. In [12], Xu addressed this over-smoothing issue and used Global variance equalization to enhance the signal when unseen noises are present. Tu [13] proposed to use a pre-emphasis filter on the time-domain signal to highlight the higher frequency details. While both approaches boost the PESQ score, the distortion are still persistent in heavily contaminated audio samples. In Fig. 4.2, we showcased some results in [17]. The RNN network [25] has 3 hidden layers of size 500 and a context window of length 3. The RNN system while extremely efficient at suppressing noise, sacrifices a great deal of the details in the spectrum, especially in lower frequency range. EHNet's results are much better and are more detailed in both the higher and lower frequency ranges. The trade-off is that EHNet distorts the speech signal by over-smoothing noisy speech. Both EHNet and our proposed ResCRN introduce distortions in the higher frequency range, where noise power overshadows speech power. Although ResCRN preserves better speech characteristics, the overall speech quality is perceived to be noisier. In Figures 3.2 - 3.4, we see that some unvoiced segments have significant residual noises as opposed to EHNet results shown here.



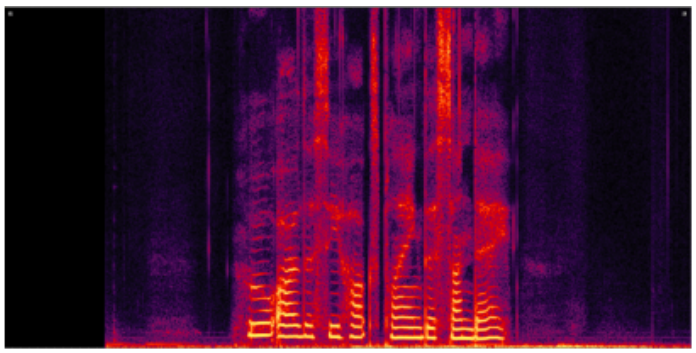
(a) Noisy Spectrogram



(b) EHNet Enhanced



(c) RNN Enhanced



(d) Clean Spectrogram

Figure 4.2: Figures from the paper Convolutional-Recurrent Neural Networks for speech enhancement [17]

4.3 Training Data Mixture

Xu [12] pointed out that a large and non-redundant training data set is crucial to the performance of DNN. For their settings, they mixed the whole TIMIT clean training data with 104 noisy data samples. SNR ratios ranged from -5 to 20dB with six different values, resulting in 2500 hours of mixture data. The performance of the network increased linearly with the size of training data and saturated around 100 hours. Since TIMIT training data has only 4 hours of clean speech, redundancy in the training set caused the performance to degrade with more data. Naturally the performance also increased when more distinct types of noises were included. However, the network trained with small dataset and using 4 noise types was already able to outperform logMMSE methods, demonstrating that deep neural networks are indeed quite powerful for speech enhancement tasks.

In the early stages of the research, we used Audible audiobooks as training data, including three male readers and three female readers and a total of 110 hours of clean, non-redundant speech data. The author believed that the rich lexical context of the audiobooks would also facilitate the neural network to learn the complex mapping function between noisy and clean speech. The network is trained on 12 hours of mini data set and again on 110 hours of whole data set, and did not show much improvement in respect to our evaluations matrices. Hence, it is important to have a diverse collection of speakers with rich lexical contents to improve the performance of neural network.

CHAPTER 5
CONCLUSION

This thesis paper proposed an end-to-end ResNet-like architecture (ResSE) for speech enhancement and further improved the system with LSTM layers (ResCRN). The ResSE network has a rich structure and adequate depth for exploiting spatial features in the spectrum. However, the lack of temporal modeling limits its performance at suppressing non-stationary noise. Therefore, we implemented LSTM layers in ResCRN to model the correlations of neighboring frames. The results with the addition of recurrent neural network improved the speech enhancement performance significantly, even with limited feature extraction in the frequency domain. Our experiments conclude that for speech enhancement, and especially in our sequence-to-sequence regression approach, combining both local and global speech features is very important. The experimental results also demonstrate that ResCRN is able to improve speech perceptual qualities as well as intelligibility even with unseen speaker and unseen noise scenarios. A comparison of our de-noised spectrum with some of the other RNN and CNN shows that while our system is able to reconstruct the speech signals nicely, the de-noising capability could be further improved.

For future works, we wish to better leverage the trade-off between speech distortion and de-noising capabilities. Our ResNet approach could be extended to an encoder-decoder system for speech enhancement, where the residual connection are made between parallel encoders and decoder to ensure the information flow.

REFERENCES

[1] P. C. Loizou, *SPEECH ENHANCEMENT SECOND EDITION*. CRC Press, 2013.

[2] A. E. Weiss M. and P. T., *Study and development of the INTEL technique for improving speech intelligibility*. Technical Report NSC-FR/4023, 1974.

[3] S. Boll, "Supression of acoustic noise in speech using spectral subtraction," IEEE Transaction Acoustic Speech Signal Processing, ASSP-27(2), 1979, pp. 113–120.

[4] J Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," IEEE Proceedings,67(12),1586-1604, 1979.

[5] ——, "All pole modeling of degraded speech," IEEE Transaction Acoustic Speech Signal Processing ASSP-26(3), 1978, pp. 197–210.

[6] Y. Ephraim and D. Malah, "Speech Enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Transaction Acoustic Speech Signal Processing ASSP-32(6), 1984, pp. 1109–1121.

[7] A. B. K. Paliwal, "A Speech Enhancement method based on Kalman Filtering," IEEE International Conference on Acoustics, Speech, and Signal Processing, 1987.

[8] B. S. Dendrinos M. and G Carayannis, *Speech Enhancement from noise; A regenerative approach*. Speech Community 10, 1991, pp. 45–57.

[9] Y. Ephraim and H. Van Trees, "A signal subspace approach for speech enhancement," IEEE Proceedings, 1993, pp. 355–358.

[10] P. S. Kevin W Wilson Bhiksha Raj and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," ICASSP, 2008, pp. 4029–4032.

[11] L.-R. D. C.-H. L. Yong Xu Jun Du, "An Experimental Study on Speech Enhancement Based on Deep Neural Networks," IEEE Signal Processing Letters (Volume: 21 , Issue: 1 , Jan. 2014), 2013, pp. 65–68.

[12] ——, " A Regression Approach to Speech Enhancement Based on Deep Neural Networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, pp. 7–19.

[13] X. Z. Ming Tu, "Speech Enhancement based on deep neural networks with skip connections," ICASSP, 2017, pp. 5565–5569.

[14] K. H. X. Z. S. R. J. Sun, *Deep Residual Learning for Image Recognition*. arXiv:1512.03385 [cs.CV], 2015.

[15] J. W. L. Se Rim Park, *A Fully Convolutional Neural Network For Speech Enhancement*. arXiv:1609.07132v1 [cs.LG], 2016.

[16] D. W. Ke Tan, "A Convolutional Recurrent Neural Network for Real-time Speech Enhancement," Interspeech DOI: 10.21437/, 2018, pp. 3229–3233.

[17] I. T. C.-h. L. Han Zhao Shuayb Zarar, "A Convolutional Recurrent Neural Network for Speech Enhancement," ICASSP 2018, 2018.

[18] Y. C. D. M. Kai Zhang Wangmeng Zuo and L. Zhang, "Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising," IEEE Transactions on Image Processing, 2016, pp. 3142 –3155.

[19] J. Garofolo, *Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database,* NIST Tech. Rep., 1988.

[20] G. Hu, *100 nonspeech environmental sounds*. http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech

[21] S. Bible. http://soundbible.com/.

[22] R. ITU-T, *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs International Telecommunication Union-Telecommunication Standarization Sector,* 2001.

[23] R. H. C. H. Taal R.c. Hendriks and J. Jensen, "An algorithm for intelligibility prediction of time frequency weighted noisy speech," pp. 2125–2146, 2011.

[24] R. G. Emmanuel Vincent and C. Fevotte, "Performance measurement in blind audio source separation," IEEE transactions on audio, speech, and language procesing, vol. 14, no. 4, 2006, pp. 1462–1469.

[25] T. M. O. O. V. P. N. A. Y. N. Andrew L. Maas Quoc V. Le, "Recurrent Neural Networks for Noise Reduction in Robust ASR," Thirteenth Annual Conference of the International Speech Communication Association, 2012.